A GIBBS SAMPLER ON THE N-SIMPLEX

AARON SMITH

1. Abstract

We determine the mixing time of a simple Gibbs sampler on the unit simplex, confirming a conjecture of D. Aldous. The upper bound is based on a two-step coupling, where the first step is a simple contraction argument and the second step is a non-Markovian coupling. We also present a MCMC-based perfect sampling algorithm that is based on our proof and which can be applied to Gibbs samplers that are harder to analyze.

2. Introduction

Given a measure μ on a convex body $K \subset \mathbb{R}^n$, how can we efficiently obtain independent samples from the distribution of μ? This problem arises in the computational sciences, and a frequently-used tool is Markov chain Monte Carlo (MCMC) [6]. Because MCMC methods produce nearly-independent samples only after a lengthy mixing period, a long-standing mathematical question is to analyze the mixing times of the MCMC algorithms which are in common use.

The analysis of discrete MCMC algorithms is very advanced, with precise bounds for many difficult problems as well as some general theory [9] [2]. For continuous samplers, there has been some general theory based on geometric or coupling arguments [11] [10] [16] [15], but many of the techniques built for discrete chains seem to run into technical difficulties. There are also very few well-understood simple chains, in stark contrast to the discrete theory, which has been built on many detailed analyses of specific chains (though see [13] [14] for some very nice analyses of some slower walks on the simplex). This paper is an attempt to carefully analyze a simple continuous chain, a Gibbs sampler on the n-simplex, and to illustrate the use of two powerful techniques from the discrete case, non-Markovian couplings [8] [4] [5] and coupling from the past [12].

Throughout most of this paper, we will be concerned with a Gibbs sampler X_t having the uniform distribution on the n-simplex $\Delta_n = \{X | \sum_{i=1}^n X_i = 1; X_i \geq 0\}$ as its stationary distribution. To take a move in this Markov chain, choose $1 \leq i < j \leq n$

Date: August 1, 2011.

and $\lambda \in [0,1]$ uniformly, then set $X_{t+1}[i] = \lambda(X_t[i] + X_t[j])$, $X_{t+1}[j] = (1-\lambda)(X_t[i] + X_t[j])$ and $X_{t+1}[k] = X_t[k]$ for $k \neq i, j$. This sampler was first mentioned in [2], where the mixing time was shown to be $O(n^2 \log(n))$. D. Aldous suggested [1] that the correct mixing time was $O(n \log(n))$, and we confirm this:

Theorem 1 (Simplex Mixing Time)**.** *If K_n^t is the t-step transition kernel for the Gibbs sampler described above, and U_n is the uniform distribution on Δ_n, then for $t > 7(C + 6.5)n \log(n)$ and $C > 0$,*

$$||K^t(x,y) - U(x)|| < 8n^{-C}$$

for all $x \in \Delta_n$.

After proving this lemma, we will briefly discuss how to turn the proof into a perfect sampling algorithm, and mention directions for further work.

3. Notation, Basic Lemmas and Strategy

This proof relies on the following standard lemma (see [9]):

Lemma 2 (Fundamental Coupling Lemma)**.** *If X_t, Y_t are two coupled instances of the same Markov chain, Y_0 is distributed according to the stationary distribution of the Markov chain, and τ is the first time at which $X_t = Y_t$, then $d_{TV}(X_t, Y_t) \leq P[t > \tau]$*

Throughout this note, we are interested in two Markov chains, X_t and Y_t, where X_t starts according to some distribution of our choosing and Y_t starts out uniformly over the simplex. We will describe a joint evolution of our two chains X_t and Y_t, such that at a specific time, the probability of having coupled is very high. The method for proving this is slightly unusual. In most coupling proofs, including the non-Markovian coupling in [8], there is an attempt to make the two chains get closer throughout the process. In our method, we attempt to couple only at a specific final time, and include many moves that are likely to increase the distance between the chains by a large amount. In fact, our joint distribution will generally assign 0 probability to coupling at any prior time.
Throughout the document, I try to be consistent about notation. Here is a list of some commonly used variables that I have tried to reserve, for reference while reading:

X_t: The Markov chain of interest.

Y_t: Another instance of the Markov chain, started at stationarity.

$P(t)$: A partition of $[n]$.

S**:** A piece of a partition.

i,j: Coordinates we update.

λ**,** λ_x**,** λ_y**:** Uniform random variable used to update a chain, or chains X_t and Y_t.

$w(S, x)$**:** The weight assigned by vector x to a subset $S \subset [n]$.

b, c, d, e**:** Exponents used to bound the size of certain often-used quantities. b is used to bound $\inf_i Y_t[i]$, c is used as a placeholder for comparing the scale of two quantities, d is used to bound $E[||X_t - Y_t||_2^2]$, and e is used to bound $\sup_i |X_t[i] - Y_t[i]|$.

In order to develop our global joint coupling, we describe two possible one-step couplings of X_t and Y_t. These are the 'proportional' coupling and the 'subset' coupling. Throughout, we will always choose to update entries at the same coordinates i, j in both X_t and Y_t at every step; only the uniform variable λ sometimes differs. Because of this, we often use these couplings to refer to actual one-step couplings as well as couplings of λ conditioned on the choice of update coordinates. Before defining the couplings, we need the following (standard) technical lemma.

Lemma 3 (Coupling Existence)**.** *Let $f(t)$ $g(t)$ be two probability density functions on $[0, 1]$, and let $0 < \alpha < 1$ be a real number s.t. $\alpha g(t) \leq f(t)$ for all t. Then it is possible to define random variables (X, Y) s.t. $X = Y$ with probability at least α and X is marginally distributed according to $f(t)$ while Y is marginally distributed according to $g(t)$.*

Proof: Let $r(t) = \frac{1}{1-\alpha}(f(t) - \alpha g(t)) \geq 0$; we note that $\int_0^t r(t) = \frac{1}{1-\alpha}(\int_t f(t) - \alpha \int_t g(t)) = 1$, and so $r(t)$ is the density of a distribution. Now we choose Y according to the distribution $g(t)$, and we choose $X = Y$ with probability α, and choose X according to the distribution $r(t)$ with probability $1 - \alpha$. It is clear that $X = Y$ with probability at least α, and that Y has the correct distribution. To see that X has the correct distribution, note that for any set A, $P[X \in A] = \alpha \int_A g(t) + (1 - \alpha) \int_A \frac{1}{1-\alpha}(f(t) - \alpha g(t)) = \int_A f(t)$, as we wanted.

In the proportional coupling, we choose an i, j and λ for Y_t, and then use the same choices for X_t, so that e.g. entry i in Y_t is updated to $\lambda(Y_t[i] + Y_t[j])$ while entry i in X_t is updated to $\lambda(X_t[i] + X_t[j])$. To define the 'subset' coupling, we first define the weight $w(S, x)$ that a vector x gives to a subset $S \subset [n]$ to be $w(S, x) = \sum_{s \in S} x_s$. The subset coupling of X_t and Y_t with respect to a subset S is a coupling that maximizes the probability that $w(S, X_{t+1}) = w(S, Y_{t+1})$, and we say that a given step of a subset

coupling succeeds if that equality holds at time $t+1$, and fails otherwise. Throughout, we generally care about whether such a coupling succeeds, not what happens when it succeeds or fails. For that reason, we don't pay attention to which optimal coupling we use, nor do we pay attention to what happens upon failure.
For a pair of points (x, y) in the simplex, a pair of update entries (i, j), and a subset $S \subset [n]$ of interest such that $i \in S$ and j not in S, we define $p(x, y, i, j, S)$ to be the subset coupling probability. Below, we give a lower bound for the success probability, which will be enough for us. Note that for some choices of x, y, i, j, S, the probability of success is 0 under any coupling.

Lemma 4 (Subset Coupling)**.** *For a pair of vectors (x, y) satisfying $\sup_i |x_i - y_i| \leq n^{-e}$ and $\inf_i x_i, \inf_i y_i \geq n^{-b}$, for $e > b$, we have for all sufficiently large n that $p(x, y, i, j, S) \geq 1 - 2n^{b+1-e}$ uniformly in S and possible i, j.*

Proof: We consider two elements of the simplex, $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$. Assume we are interested in coupling the weights of subset $S = I \cup (1) \subset [n]$, and that coordinates 1 and j, neither in I, are being updated. We are updating vectors x, y with the random variables λ_x, λ_y and would like to find the probability that they can be coupled. We note that the balance condition is exactly

$$\lambda_x(x_1 + x_j) + \sum_{i \in I} x_i = \lambda_y(y_1 + y_j) + \sum_{i \in I} y_i \tag{1}$$

which we can restate as

$$\lambda_x = \lambda_y \frac{y_1 + y_j}{x_1 + x_j} + \frac{1}{x_1 + x_j} \sum_{i \in I} (y_i - x_i) \tag{2}$$

Assume for now that $\frac{y_1+y_j}{x_1+x_j} > 1$. Then, by the coupling existence lemma, a valid coupling is to choose λ_y according to the uniform distribution on $(0, 1)$, and to choose $\lambda_x = \lambda_y \frac{y_1+y_j}{x_1+x_j} + \frac{1}{x_1+x_j} \sum_{i \in I} (y_i - x_i)$, as long as that is a value between 0 and 1, and to choose from some (unimportant) remainder distribution with the remaining probability. Let $m = \frac{y_1+y_j}{x_1+x_j}$ and let $\delta = \frac{1}{x_1+x_j} \sum_{i \in I} (y_i - x_i)$. Then integration tells us that this results in a successful coupling with probability $\inf(1, \frac{1-\delta}{m}) - \sup(0, -\frac{\delta}{m})$. If $\frac{y_1+y_j}{x_1+x_j} < 1$, then we choose λ_x first and then choose λ_y according to the equivalent formula; this succeeds with probability $\inf(1, 1 + \delta) - \sup(0, \delta)$. Thus, if $|m - 1|$, $|m^{-1} - 1| < \epsilon_1$ and $|\delta| < \epsilon_2$, then the coupling probability is at least $1 - 2\frac{\epsilon_2}{1-\epsilon_1}$.
Next, note that if $\sup_i |x_i - y_i| \leq n^{-e}$ and $\inf_i x_i, \inf_i y_i \geq n^{-b}$, then we can take $\epsilon_1 = n^{b-e}$ and $\epsilon_2 = n^{b+1-e}$. This proves the lemma.

We can now describe the overall strategy. In the first phase, we show that the proportional coupling is contractive, and prove that the two chains are very close with high probability after about $n \log(n)$ steps. In the next phase, we run Y_t until a

specified time T. We use information about Y_t from time 0 to T to construct a nested sequence of partitions, starting with one set at time 0 and normally ending with n singletons. We then show that, if the partitions are chosen correctly, it is possible to make sure that each piece of the partition has the same weight in both X_t and Y_t for all times, with high probability. When the final partition consists of only singletons, this implies that the chains have coupled. The main difficulties here are constructing the partition, and then showing that the conditions required for the subset coupling lemma apply with high probability.

It is worth pointing out that the dependence of the coupling on the future is in fact necessary to get the correct mixing time, or indeed any bound that is $o(n^2)$. This is analogous to the well-known fact that no Markovian coupling of the random transposition walk on S_n can give a coupling time that is $o(n^2)$.

4. First Coupling Stage

We define $Z_t = ||X_t - Y_t||_2^2$. We then compute $E[Z_t|X_{t-1}, Y_{t-1}]$ under the 'proportional coupling' described above. We have

Lemma 5 (Burn-In)**.** *After $t = \frac{3}{2}dn\log(n)$ steps of the proportional coupling,$E[Z_t] \leq 2n^{-d}$.*

Proof: We perform the following computation

$$\begin{aligned}
E[Z_t|X_{t-1},Y_{t-1}] &= E[\frac{1}{n(n-1)}\sum_{i\neq j}[\sum_{k\neq i,j}(x_k-y_k)^2+2\lambda^2(x_i+x_j-y_i-y_j)]] \\
&= \frac{1}{n(n-1)}\sum_{i\neq j}[\sum_{k\neq i,j}(x_k-y_k)^2+\frac{2}{3}(x_i-y_i)^2 \\
&+\frac{2}{3}(x_j-y_j)^2+\frac{4}{3}(x_i-y_i)(x_j-y_j)] \\
&= (1-\frac{2}{3(n-1)}+\frac{2}{3n(n-1)})Z_{t-1} \\
&+\frac{4}{3n(n-1)}\sum_{i\neq j}(x_i-y_i)(x_j-y_j) \\
&= (1-\frac{2}{3(n-1)}+\frac{2}{3n(n-1)})Z_{t-1}+\frac{4}{3n(n-1)}((\sum_i(x_i-y_i))^2 \\
&-\sum(x_i-y_i)^2) \\
&= (1-\frac{2}{3(n-1)}+\frac{2}{3n(n-1)})Z_{t-1}+0-\frac{4}{3n(n-1)}Z_{t-1} \\
&= (1-\frac{2}{3(n-1)}-\frac{2}{3n(n-1)})Z_{t-1}
\end{aligned}$$

And so in particular, $E[Z_t] < (1-\frac{2}{3n})^t Z_0 < 2(1-\frac{2}{3n})^t$, and so at time $S = \frac{3}{2}dn\log(n)$, $E[Z_t] \leq 2n^{-d}$, proving the lemma.

By Markov's inequality, $P[|X_t[i]-Y_t[j]| > \epsilon] \leq \epsilon^{-1}n^{-\frac{1}{2}d}$. As an aside, if we couple X_t and Y_t by choosing the same coordinates i, j and uniform random variables λ, λ', one sees that the one-step coupling which minimizes the distance Z_t described above occurs when $E[\lambda\lambda']$ is maximized, and this maximum occurs at $\lambda = \lambda'$.

5. Second Coupling Stage

Assume that at time 0, X_0 and Y_0 satisfy $||X_0-Y_0||_2^2 \leq n^{-d}$. We describe a coupling from time 0 to time $T = (\frac{1}{2}+\epsilon)n\log(n)$, which has a high chance of succeeding for any $\epsilon > 0$. First, we choose a sequence of pairs of distinct elements $(i(t), j(t))$ for time $1 \leq t \leq T$. From this sequence, we will construct a nested sequence of partitions of $[n]$, $P(0) \geq P(1) \geq \ldots \geq P(T)$, where we say that partition A is less than partition B if every piece of partition A is contained in a piece of partition B. We will also construct a sequence of marked times $t_1, \ldots, t_k$, and a sequence of graphs G_t on the vertex set $[n]$.

The first partition, $P(T)$, consists of n singletons, our list of marked times is initially

empty, and the graph G_T has no edges. To construct $P(t-1)$ from $P(t)$, we look at the edge $(i(t-1), j(t-1))$. If it goes between elements in two parts of $P(t)$, we join those two parts to make $P(t-1)$; otherwise, we set $P(t-1) = P(t)$. If two parts are joined at time $t-1$, we call them $S(t,1)$ and $S(t,2)$, and add time t to our list of marked times. Finally, at time t we add edge $(i(t-1), j(t-1))$ to the graph G_t if it wasn't already present, to create G_{t-1}. We make a few elementary observations. The first is that there are at most $n-1$ marked times. The next is that there are $n-1$ marked times if and only if $P(0) = [n]$ which in turn occurs if and only if G_0 is connected.
The question of whether G_t is connected is classical. The following result, found [3] among other places, is good enough for our purposes:

Lemma 6 (Connectedness)**.** *For $T > (\frac{1}{2} + \epsilon)n\log(n)$ and $\epsilon > 0$, the probability that G_0 is greater than $1 - 2n^{-\epsilon}$ for n sufficiently large.*

Having constructed this partition, we now couple X_t and Y_t from time $0 \le t \le T$. First, we need to choose the coordinates to update; we do this by updating coordinates $i(T-t)$ and $j(T-t)$ at time t in both chains. Note that this is a time reversal with respect to the order we looked at coordinates when building the graph. Since our chain is reversible, this doesn't present a problem for our coupling. Next, we must describe the coupling of the coordinates. If $T-t$ is a marked time, then we perform a subset coupling for the subset $S(T-t,1)$. Otherwise, we do a proportional coupling. I claim that this couples the two walks by time T with high probability. To prove this, we need the following three lemmas.

Lemma 7 (Technical Lemma 1)**.** $\prod_{j=1}^{n}(1 + kj^{-1}) \le (k+1)n^k$

Proof: This is a straightforward induction on n. Note that the statement is true for $n = 1$ and all k, and assume it is true for $1 \le n \le N$. Then we note that

$$\begin{aligned}\prod_{j=1}^{N+1}(1+kj^{-1}) &= (1+\frac{k}{N+1})\prod_{j=1}^{N}(1+kj^{-1}) \\ &\le (1+\frac{k}{N+1})(k+1)N^k \\ &< (k+1)N^k + (k^2+k)N \\ &< (k+1)(N+1)^k\end{aligned}$$

Proving the lemma.

Lemma 8 (Technical Lemma 2)**.** *Let $n = S_0 > S_1 > \ldots > S_k = m$ be a decreasing sequence of integers, such that $S_{j+1} \ge \frac{1}{2}S_j$. For such a sequence, define $f(S_0, \ldots, S_k) = \prod_{j=0}^{k-1}(1 + \frac{S_j - S_{j+1}}{S_j})$, and let $F(n,m)$ be the supremum of $f(S_0, \ldots, S_k)$ over all such sequences. Then $F(n,m) = f(n, n-1, \ldots, m)$. Further, $F(n,m) \le 2n$*

Proof: We prove this by simultaneous induction on n, m. It is clear for all n that $F(n, n-1) = f(n, n-1)$. Now fix an N and M, and assume that for all $m > M$ $F(N, m) = f(N, N-1, \ldots, m)$ and that for all $n < N$ and all m, $f(n, m) = f(n, n-1, \ldots, m)$. We will show that $F(N, M) = f(N, N-1, \ldots, M)$. To do so, assume this isn't the case. Then there is some other sequence $N = S_0, \ldots, S_k = M$ such that $f(S_0, \ldots, S_k) > f(N, \ldots, M)$. From the induction hypothesis, this is impossible if $S_1 = n-1$, for if we did, we would have

$$\begin{aligned} f(S_0, \ldots, S_k) &= (1 + \frac{1}{n}) f(S_1, \ldots, S_k) \qquad \leq (1 + \frac{1}{n}) f(n-1, \ldots, m) \\ &= f(n, \ldots, m) \end{aligned}$$

Thus, we can assume $S_1 < n-1$. By the induction hypothesis, we must have $S_{j+1} = S_1 - j$ for all $j \geq 1$, and so $f(S_0, \ldots, S_k) > f(N, \ldots, M)$ implies

$$\begin{aligned} (1 + \frac{S_1}{n}) &> \prod_{j=1}^{n-S_1} (1 + \frac{1}{n-j+1}) \\ &\geq (1 + \frac{S_1 - 1}{n})(1 + \frac{1}{n - S_1 + 1}) \\ &> 1 + \frac{S_1}{n} \end{aligned}$$

Where the second line is due again to the induction hypothesis. This is a contradiction, and we conclude $F(n, m) = f(n, n-1, \ldots, m)$ for all n, m.
To prove the last part, we note that

$$\begin{aligned} F(n, m) &\leq F(n, 1) \\ &= \prod_{j=1}^{n} (1 + j^{-1}) \\ &\leq 2n \end{aligned}$$

where the last step is due to technical lemma 1, above.

Lemma 9 (Largeness)**.** $P[\inf_{1 \leq i \leq n} \inf_{0 \leq t \leq T} Y_t[i] \leq n^{-4.5}] = o(1)$

Proof: This calculation is done on pp. 11 of chapter 13 of **[AF]**.

Lemma 10 (Smallness)**.** *Let $A(s)$ be the set on which $\sup_i |X_t[i] - Y_t[i]| \leq 2n^e$ and $\inf_i x_i, \inf_i y_i \geq n^{-4.5}$ for all $0 \leq t \leq s$, and call this 'condition A'. Then $P[(\sup_{0 \leq s \leq t} ||X_s - Y_s||_2^2 \geq n^c ||X_0 - Y_0||_2^2) \cap A(t-1)] \leq 2tn^{2-c}$ for all $c > 0$.*

Proof: We will estimate $E[||X_t - Y_t||_2^2]$ while conditioned on something slightly stronger than all subset couplings succeeding until time t, by bounding how much the term changes during each successful 'subset coupling'. At the same time, we will

estimate the probability that condition A fails at time t. To set notation, assume the subset couplings happen at times $t_1, \dots, t_{n-1}$ between sets $S(t_k, 1)$ and $S(t_k, 2)$ where $|S(t_k, 1)| \leq |S(t_k, 2)|$. We write $||X_t - Y_t||^2_{2,S} = \sum_{s \in S}(X_t[s] - Y_t[s])^2$, the L^2 norm restricted to indices contained in the subset S. We let $i(t_k) \in S(t_k, 1)$ and $j(t_k) \in S(t_k, 2)$ be the two coordinates updated at time t_k. If the coupling at time t_k was successful, then we have $X_{t_k}[i(t_k)] = Y_{t_k}[i_{t_k}] + \sum_{l \in S(t_k,1)}(Y_{t_k}[l] - X_{t_k}[l])$. In particular, we can make the following computation. In it, the expectation is taken with X_{t_k-1}, Y_{t_k-1}, t_k, the size of $S(t_k, 1)$ (but not the set), the old partition $P(t_k)$ (but not the new partition $P(t_k + 1)$) and the part of the partition p that is being broken in this step fixed; we are viewing the particular sets $S(t_k, 1)$ and $S(t_k, 2)$, and the coordinates $i(t_k)$ and $j(t_k)$, as the random variables. We also condition on the update variable λ_y at time t being in the range described in the 'subset coupling lemma' based on condition A (looking at that lemma shows that this is just conditioning on $\alpha < \lambda_y < \beta$ for some fixed $\alpha(n) \sim 0$ and $\beta(n) \sim 1$). Note that we can condition on λ_y being in that range even if condition A doesn't hold; we will simply no longer be guaranteeing that the subset couplings all succeed. As an aside, I condition on λ_y being in a particular range rather than on the subset coupling succeeding because the partition process $P(t)$ is independent of the update variables λ_y, but it is not independent of the event that a given subset coupling succeeds.
I will write $F(t_k)$ for the σ-algebra described above. The reason for using this σ-algebra is that, conditioned on $P(t_k)$, the part of the partition $p(t_k)$ to be split int two at time t_k, and the size of $S(t_k, 1)$, all subsets of p of size $|S(t_k, 1)|$ are equally likely. Thus, we find that if condition A holds at time t_k,

$$\begin{aligned}
E[(X_{t_k}[i(t_k)] - Y_{t_k}[i(t_k)])^2 | F(t_k)] &= E[(\sum_{l \in S(t_k,1)} (Y_{t_k-1}[l] - X_{t_k-1}[l]))^2] \\
&= E[\sum_{l \in S(t_k,1)} (Y_{t_k-1}[l] - X_{t_k-1}[l])^2] + E[\sum_{l \neq m \in S(t_k,1)} (Y_{t_k-1}[l] - X_{t_k-1}[l])(Y_{t_k-1}[m] - X_{t_k-1}[m])] \\
&\leq \frac{|S(t_k, 1)|}{|p|} ||X_{t_k-1} - Y_{t_k-1}||^2_{2,p} - \frac{|S(t_k, 1)|(|S(t_k, 1)| - 1)}{|p|(|p| - 1)} ||X_{t_k-1} - Y_{t_k-1}||^2_{2,p} \\
&= \frac{|S(t_k, 1)|}{|p|}(1 - \frac{|S(t_k, 1)| - 1}{|p| - 1}) ||X_{t_k-1} - Y_{t_k-1}||^2_{2,P}
\end{aligned}$$

Where in the third line I use the fact that $\sum_{s \in S(t_k,1) \cup S(t_k,2)}(X_{t_k-1}[s] - Y_{t_k-1}[s]) = 0$ by the assumption that all subset couplings up to this time have succeeded. This implies that

$$E[||X_{t_k} - Y_{t_k}||^2_{2,p} \, | F(t_k)] \leq (1 + \frac{|S(t_k, 1)|}{|p|}(1 - \frac{|S(t_k, 1)| - 1}{|p| - 1})) ||X_{t_k-1} - Y_{t_k-1}||^2_{2,p}$$

and also that

$$E[||X_{t_k} - Y_{t_k}||_2^2 \ |F(t_k)] \leq (1 + \frac{|S(t_k,1)|}{|p|}(1 - \frac{|S(t_k,1)|-1}{|p|-1}))||X_{t_k-1} - Y_{t_k-1}||_2^2$$

Before the next step, we note that the same argument which shows that the proportional coupling is contractive on the L^2 distance between X_t and Y_t shows that it is contractive on the L^2 distance between X_t and Y_t restricted to any part $p(t)$ of the partition $P(t)$, without any change. Because of this, in the following calculations we are free to ignore the proportional coupling steps; following along, it is easy to see that they would only improve our estimates. We also note that, for any 'part' of a partition, only one entry is ever used to 'link' it when creating the next-coarser partition. Let $I_{i,k}$ be the indicator function of i being in $p(t_k)$, and $J(i) \subset (t_1, \ldots, t_n)$ be the set of marked times where the splitting partition contains index i. Then we find that, for any particular i, and for $t_{k+1} > t > t_k$,

$$\begin{aligned}
&E[(X_t[i] - Y_t[i])^2; A(t-1)] = E[[E[\ldots E[(X_T[i] - Y_T[i])^2 | F(t_k)] | F(t_{k-2})] \ldots | F(t_1)]; A(t-1)] \\
&\leq E[E[\ldots E[(1 + \frac{|S(t_k,1)|}{|p(t_k)|} I_{i,k})||X_t - Y_t||^2_{2,p(t_k)}] \ldots | F(t_1)]; A(t-1)] \\
&\leq E[\prod_{s \in J(i) \cap (0,t]} (1 + \frac{|S(s,1)|}{|p(s)|}); A(t)]||X_0 - Y_0||_2^2 \\
&\leq E[2n||X_0 - Y_0||_2^2] \\
&= 2n||X_0 - Y_0||_2^2
\end{aligned}$$

Where the second last line is by technical lemma 2. There is a question as to why we can take this calculation over the set where condition A holds. The answer is that at marked times we pretend we are running the chain by choosing λ_x, λ_y according to the relation (2); this chain, of course, is no longer confined to being in the simplex. This preserves the formula used above, and at the end we just set to 0 all contributions from parts of the chain which left the simplex, which can only decrease the expectation. Continuing, we sum this over all i, to find that

$$E[||X_t - Y_t||_2^2; A(t)] \leq 2n^2||X_0 - Y_0||_2^2$$

Thus, by Markov's inequality, and taking a union bound over $0 \leq s \leq t$, we have

$$P[(\sup_{0 \leq s \leq t} ||X_s - Y_s||_2^2 \geq n^c||X_0 - Y_0||_2^2) \cap A(t)] \leq 2tn^{2-c}$$

where we implicitly use the fact that $P[(X > \epsilon) \cap S] = P[1_S X > \epsilon]$ for nonnegative random variables X and $\epsilon > 0$.

Lemma 11 (Condition Lemma)**.** *Condition A as defined in the smallness lemma holds until time T with probability at least $1 - 2n^{5+2e-d} + o(1)$.*

Proof: From the smallness lemma and the proof of the largeness lemma, we find for all s that

$$\begin{aligned} P[A(s)^c] &\leq P[\sup_{t\leq s} ||X_t - Y_t||_2^2 \geq n^{-2e}] + o(\frac{1}{n}) \\ &\leq 2sn^{2-2e-d} + P[A(s-1)^c] + o(\frac{1}{n}) \\ &\leq 2n^{5+2e-d} + P[A(0)^c] + o(1) \\ &= 2n^{5+2e-d} + o(1) \end{aligned}$$

which proves the lemma. We note that this is one of several places that an extra factor of n is picked up; this turns out to be fairly unimportant.

Lemma 12 (Weight Lemma)**.** *Assume* $||X_0 - Y_0|| \leq n^{-d}$*. Then the equality*

$$w(S, X_t) = w(S, Y_t) \tag{3}$$

holds for all $0 \leq t \leq T$ *and all* $S \subset P(t)$ *with probability at least* $1 - n^{5+2e-d} - n^{6.5-e} + o(1)$*, for any choice of* $e > 0$*.*

Proof: The equality clearly holds at time 0, and it can only become false at a marked time, since at unmarked times t, the weights of parts of $P(t)$ cannot change in either X_t or Y_t. So, assume that the proposition is true for time $t < t_k$. Then note that if the subset coupling is successful at time $t = t_k$, $w(S(t,1), X_t) = w(S(t,1), Y_t)$ by construction. However, by induction, $w(S(t-1,1) \cup S(t-1,2), X_{t-1}) = w(S(t-1,1) \cup S(t-1,2), Y_{t-1})$, and so $w(S(t,2), X_t) = w(S(t,2), Y_t)$ as well. Since none of the other parts of $P(t-1)$ change weight, this implies that the proposition remains true unless one of the n subset couplings fails.
If condition A holds, the probability of any coupling being successful is at least $1 - n^{5.5-e}$ by the subset coupling lemma. By the condition lemma, this condition holds for all n with probability at least $1 - n^{5+2e-d} + o(1)$. Thus, a union bound tells us that the probability of any of the n subset couplings failing is less than $n^{5+2e-d} + n^{6.5-e} + o(1)$, proving the lemma.

Now recall that if the graph G_T is connected and all components $K \in P(t)$ satisfy $w(K, X_t) = w(K, Y_t)$ for $0 \leq t \leq T$, then at time T the two walks have coupled. Assume that we have a burn-in period of $6Cn\log(n)$ and a final coupling period of $T = Cn\log(n)$. Then choosing $d = 3C$ and $e = C$ in the burn-in lemma gives us a total chance of coupling greater than $1 - 2n^{-C} + n^{5-C} + n^{6.5-C} + 2n^{0.5-C}$. This proves our theorem. Note that a different choice of d, e for C small gives us some 'pre-cutoff' at a slightly smaller overall multiple of $n\log(n)$ steps.

6. Lower Bound

Since our walk is over a continuous space, coupling cannot have occurred unless each coordinate has been chosen. Since only two coordinates are chosen at a time, the classical coupon-collector results in [7] tells us that at time $T = \frac{1}{2}n(\log(n) - c)$, $d_{TV}(X_t, Y_t) \geq 1 - exp(-exp(c)) + o(1)$ as n goes to infinity.
It is possible to do a little better than this. Assume we start at position $(1, 0, \ldots, 0)$. We then try to 'collect' the n coordinates, but only count a coordinate as 'collected' if it is chosen at the same time as a non-zero coordinate. Next, let τ_j be the expected time to collect the $j'th$ coupon after collecting the $j-1$'st. We note that $E[\tau_1] = n$, $E[\tau_2] = 0$. For larger j, $E[\tau_{j+1}] = \frac{n^2}{j(n-j)}$. Thus,

$$E[\sum_{j=1}^{n} \tau_j] = n + n^2 \sum_{j=2}^{n-1} \frac{1}{j(n-j)}$$
$$\sim 2n \log(n)$$

And in particular the same concentration argument tells us that the mixing time is at least $2n(\log(n) - c(\epsilon))$.

7. Closely Related Walks

It is worth pointing out a small number of cases where the above argument goes through with very few changes. The first allows us to go from sampling from the uniform distribution to sampling from a large class of distributions on the simplex, including symmetric Dirichlet distributions. At each step of the random walk, instead of choosing λ according to the uniform distribution on $[0, 1]$, choose it according to some other distribution with twice differentiable cdf F satisfying $F[x] = 1 - F[1 - x]$ for all $0 \leq x \leq \frac{1}{2}$. Then the above arguments show that the total mixing time is $O(nlog(n)\frac{||F''||_\infty}{1-2E[\lambda^2]})$, essentially without modifcation.
It is also possible to apply this argument to its 'discrete' analogue, in which M indistinguishable balls are stored in n boxes; these are known as M-compositions of n. The analogous Markov chain involves choosing two boxes at every step, and having each ball switch with probability 1 2 . The same argument applies to the discrete chain, giving a mixing bound of order $O(nlog(n))$, but there need to be enough balls for the continuous approximation to be good at each step. A straightforward step-through of the argument gives a bound of $M > n^{18.5}$ above, or $M > n^{5.5}$ for Aldous' greedy coupling.
There is a follow-up paper being prepared which will discuss a wider variety of related walks, requiring larger modifcations.

8. Perfect Sampling on the Simplex

In this section, I discuss how the two-chain coupling described above can be modified into a grand coupling, and how to use this fact to create a perfect sampling algorithm. First, we recall the coupling from the past (CFTP) algorithm, described in greater detail in [12]. First, choose some large time T, and start a copy of the Markov chain X_{-T}^{ω} for each ω in the sample space Ω. Next, couple all of the chains from time $-T$ to time 0. If the chains have coallesced by time 0, the resulting single value is distributed according to the stationary distribution of the chain. If not, we couple chains started at all points from $-2T$ to T and keep the evolution from $-T$ to 0, then from $-3T$ to $-2T$ keeping the evolution from $-2T$ to 0, and so on until coallescence at 0 has occurred.
For Markov chains on a finite state space, it is easy in theory to construct a grand coupling that will eventually coallesce, though bad couplings are very inefficient. In practice, even on finite chains, CFTP is only used if the chain has some very special properties. The most popular property, introduced in the original paper, is 'monotonicity'. Briefly, we introduce a partial order $\leq$ on Ω, and say that a coupling of two chains X_t, Y_t is monotone if $X_0 \leq Y_0$ implies $X_t \leq Y_t$ for all $t > 0$. It is then easy to see that if our grand coupling is monotone, it is sufficient to keep track of chains started at maximal and minimal elements of the poset. If they have coupled, all states have coupled. Most uses of CFTP rely on monotonicity or its twin, antimonotonicity. For Markov chains on infinite state spaces, many grand couplings will never coallesce, and of course we can't keep track of all of the starting values on a computer. Some chains have a monotonicity property, but such a property isn't obvious for the simplex model. Despite this, there is a fairly efficient perfect sampling algorithm that requires tracking only $n + 1$ points (and a little extra overhead each time an epoch of length T fails).

Let X_t^v be a copy of the Markov chain started at v at time 0, and let e_j be the j'th unit standard basis vector. We will do a proportional coupling from time $0 < t < T_1$. Then we note that there exists a matrix $\lambda_t^i[j]$ such that for any v, $X_t^v = \sum_{j=1}^{n} \lambda_t^i[j] v_j$

. Thus,

$$\begin{aligned}
||X_t^v - X_t^w||_1 &= \sum_{i=1}^n |X_t^v[i] - X_t^w[i]| \\
&= \sum_{i=1}^n |\sum_{j=1}^n \lambda_t^i[j](v_j - w_j)| \\
&\leq \sum_{i=1}^n n \sup_j \lambda_t^i[j] \\
&= \sum_{i=1}^n \frac{n}{n-1} \sum_{j,k} |X_t^{e_j}[i] - X_t^{e_k}[i]| \\
&= (1 - \frac{1}{n}) \sum_{j,k} ||X_t^{e_j} - X_t^{e_k}||_1
\end{aligned}$$

Applying the older burn-in inequality to the expectation of $||X_t^{e_j} - X_t^{e_k}||_1$ for all distinct pairs j, k, then Markov's inequality to the probability that the above norm is large, we find that for $t > \frac{3}{2}mn\log(n)$,

$$P[||X_t^{e_j} - X_t^{e_k}||_1 > n^{-m+20}] < n^{-20}$$

and so taking a union bound, and applying the inequality proved just above,

$$P[\sup_{v,w \in \Delta_n} ||X_t^v - X_t^w||_1 > n^{-m+17}] < n^{-16}$$

This tells us that after a first step of $O(n\log(n))$ steps, the L^1 distance between any two points is extremely small. The second step of the coupling is almost identical to the proof given in the first section of this note. In particular, we run $X_t^{(n^{-1},\ldots,n^{-1})}$ from time T_1 to time T, recording all choices of edge and averaging variable. We then form the same partition process, and use it to attempt maximal couplings of all variables to this special chain. As long as the $X_t^{e_j}$ continue to be a sufficiently small ball as measured in L^1 metric, and $X_t^{(n^{-1},\ldots,n^{-1})}$ remains far from 1, we can simultaneously couple all of these chains with high probability.

It remains to determine what happens if one of the above maximal couplings fails. First of all, when continuing that run, we should immediately switch to a proportional coupling rather than continue attempting maximal couplings; this minimizes the amount of computation we will need to do later. Then, simply try again on the interval $[-2T, -T]$. If the coupling succeeds this time, we will need to determine what happens to the special chain starting at $(n^{-1}, \ldots, n^{-1})$ at time $-2T$. Fortunately, this is not so difficult. We already know what has happened to it up to time $-T$. All of the proportional coupling steps that occur between $-T$ and 0 are easy to construct,

due to linearity and the fact that we have recorded what has happened to the vertices of the simplex. Successful maximal coupling steps are similarly easy to record, since the proof of fast coupling for the simplex gives a linear equation which can be used to determine the new coordinates. The only difficulty is what happens during the single unsuccessful maximal coupling. Fortunately, the remainder measure after a failed maximal coupling is just the sum of at most two uniform measures. When we get to this point, simply sample from that mixture independently of everything that has happened until that point.

It should be noted that the above algorithm isn't special to our particular walk or target distribution on the simplex. For example, it will also work for the earlier small modifications based on changing the distribution of λ We also observe that this algorithm doesn't need an a priori bound on the mixing time to run. In fact, running the algorithm can be used to rigorously check an estimated bound of time T, simply by running and checking convergence. This has the advantage of being fairly efficient even when it is not easy to draw from the stationary distribution.

9. Acknowledgements

The author thanks David Aldous for mentioning the problem, and Persi Diaconis, Olena Bormashenko and Daniel Jerison for many helpful conversations.

References

[1] David Aldous. Open problems. http://www.stat.berkeley.edu/ aldous/Research/OP/index.html.
[2] David Aldous and Jim Fill. *Reversible Markov Chains and Random Walks on Graphs*. 1994.
[3] Bela Bollobas. *Random Graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Press Syndicate of the University of Cambridge, Cambridge, 2001.
[4] Olena Bormashenko. A coupling proof for random transpositions. *Preprint*, 2011.
[5] Robert Burton and Yevgeniy Kovchegov. Mixing times via super-fast coupling. *Preprint*, 2011.
[6] Persi Diaconis. The markov chain monte carlo revolution. *Bull. Amer. Math. Soc.*, 46(1):179–205, 2008.
[7] Paul Erdos and Alfred Renyi. On a classical problem of probability theory. *Magyar Tud. Akad. Mat. Kutato. Int. Kozl*, 1961.
[8] Tom Hayes and Eric Vigoda. A non-markovian coupling for randomly sampling colorings. *FOCS proceedings*, 2003.
[9] Yuval Levin, David; Peres and Elizabeth Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, Rhode Island, 2009.
[10] Laslo Lovasz. Hit and run mixes fast. *Math. Prog.*, 1998.
[11] Laslo Lovasz and Santosh Vempala. Hit and run is fast and fun. *Technical Report - Microsoft Research*, 2003.
[12] James Propp and David Wilson. Exact sampling with coupled markov chains and applications to statistical mechanics. *Rand. Struct. Alg.*, 9:223–252, 1996.
[13] Dana Randall and Peter Winkler. Mixing points on a circle. *Lecture Notes in Computer Science*, 3624:426–435, 2005.

[14] Dana Randall and Peter Winkler. Mixing points on an interval. *Proceedings of ANALCO*, 2005.
[15] Jeffrey Rosenthal. Minorization conditions and convergence rates for markov chain monte carlo. *JASA*, 90:558–566, 1995.
[16] Wai Kong Yuen. Applications of geometric bounds to convergence rates of of markov chains and markov processes on rn. *PhD Thesis*, 2001.

Department of Mathematics, Stanford University, Stanford, CA 94305
E-mail address: asmith3@math.stanford.edu